ThatiAR: Subjectivity Detection in Arabic News Sentences

Anonymous ACL submission

Abstract

In this study, we present the first large dataset for subjectivity detection in Arabic, consisting of ∼3.6K manually annotated sentences, and GPT-4o based explanations. In addition, we include instructions (both in English and Arabic) to facilitate LLM based fine-tuning. We provide an in-depth analysis of the dataset, annotation process, and extensive benchmark results, including PLMs and LLMs. Our analysis of the annotation process highlights that annotators were strongly influenced by their political, cultural, and religious backgrounds, especially at the beginning of the annotation process. The experimental results suggest that LLMs with in-context learning provide better performance. We release the dataset and resources to the community.[1]

1 Introduction

Detecting subjectivity[2] in news sentences is crucial for several reasons. It helps identifying media bias by distinguishing between objective reporting and subjective content, thereby enhancing the credibility of news sources. This differentiation is also vital in combating misinformation and fake news by flagging opinion-based content for further verification. In Figure 1, we present an example of a subjective sentence that can be misleading and cause fear among citizens. The highlighted part of the text in the example is subjective.

With the reliance on social media as platforms of expression, users often resort to informality, dialects, and a combination of languages. When seeking news reports and statements, readers turn to news outlets for knowledge and assessments of current events. While readers may consider news from reliable outlets as objective sources of information, research shows that news reports are often partisan, subjective, and reflective of the news agency's standpoint. Therefore, analyzing subjectivity provides insights into public sentiment and the social impact of news. It can empower readers to make informed decisions and encourages critical thinking by highlighting subjective reports.

للأسف، انتشرت قوات الشرطة بكثافة اليوم في العاصمة الجزائر، مما جعل الأجواء متوترة بينما حلقت طائرات مروحية في سماء المدينة.

Translation: Unfortunately, police forces spread densely today in the capital, Algiers, creating a tense atmosphere while helicopters flew over the city.

Figure 1: An example of a subjective sentence that can be misleading and cause fear.

While there has been research effort to develop methods and systems to automatically identify such content, the majority of studies focus on English or other high-resourced languages. However, the field is growing to incorporate "morphologically-rich" or complex languages, including Urdu, Arabic, and Turkish (Abdul-Mageed, 2015).

Research on subjectivity in Arabic content (Abdul-Mageed et al., 2011; Abdul-Mageed and Diab, 2011, 2012; Mourad and Darwish, 2013; Abdul-Mageed and Diab, 2014) addresses the complexities of language usage, primarily focusing on Modern Standard Arabic (MSA) and regional dialects. The significant variation in Arabic dialects across different geographical and national contexts presents an additional challenge. Therefore, in this study, we focused on Arabic, with a special emphasis on news content. Given the lack of resources for developing AI-based sys-

[1] anonymous.com

[2] Subjectivity, according to Abo et al. (2019), "refers to aspects of language used to express feelings, opinions, evaluations, and speculations and, as such, it incorporates sentiment".

tems in Arabic, we introduce *ThatiAR*,[3] a reasonably large and well balanced dataset consisting of manually annotated news sentences. While annotating *ThatiAR*, we addressed three research questions: characteristics of news report, annotators' perceptions, and the applicability of current annotation guidelines, with regards to subjectivity (see Section 3.3). We conducted extensive experiments to create a benchmark using different Pre-trained Language Models (PLMs) and Large Language Models (LLMs) that can serve as a foundation for future research. Given that current LLMs consistently push the boundaries of NLP and achieve state-of-the-art performance in tasks such as machine translation, summarization, sentiment analysis, and more complex applications like legal document analysis and creative writing (Liang et al., 2022; Bang et al., 2023; Ahuja et al., 2023; Hendy et al., 2023; Khondaker et al., 2023; Abdelali et al., 2024), therefore, we used GPT-4o to generate explanations for why a sentence is labeled as subjective or objective. Additionally, we developed instructions for each data point, resulting in a comprehensive instruction-following dataset. Below is a summary of our contributions:

- We developed *ThatiAR*, a dataset consisting of approximately 3.6K manually annotated news sentences. This is largest dataset compared to any other subjectivity dataset released so far.
- We provide a detailed analysis of the annotation process, addressing the research questions mentioned earlier.
- Benchmark results using different PLMs and LLMs.
- The dataset includes explanations for the provided labels, which can aid in developing explanation-based generative models.
- An instruction-following dataset that can help in building models capable of following instructions.

2 Related Work

Research on subjectivity analysis often approaches subjectivity and sentiment analysis hierarchically. First, texts are classified as subjective or objective, and then sentiments are designated as "positive," "negative," or "mixed" for the subjective texts (Korayem et al., 2012; Mourad and Darwish, 2013; Refaee and Rieser, 2014a,b). Typically, it has been served as a preliminary step to sentiment analysis (Savinova and Moscoso Del Prado, 2023), as it relies primarily on subjective fragments of the text. Earlier approaches of research for this domain was mainly rule based and mostly for English. Recently the problem has been mostly addressed by training transformer based models (Huo and Iwaihara, 2020).

For Arabic, earlier research by Abdul-Mageed et al. (2014) proposed a system for sentence-level subjectivity analysis of Arabic social media. They also developed a comprehensive corpus that includes sentences from chats, tweets, Wikipedia pages, and web forums, which were manually annotated as objective, subjective, neutral, or mixed, and further categorized by sentiment (i.e., positive and negative). Habash et al. (2013) developed the Qatar Arabic Language Bank (QALB), which provides guidelines for Arabic corpus annotations that account for the Qatari dialect. These corpora and dataset developments are significant to the field of subjectivity and sentiment analysis in both Modern Standard Arabic (MSA) and Dialectal Arabic (DA). Additionally, Azmi and Alzanin (2014) developed an opinion mining system targeting the Saudi Najdi Dialect, called Ara'a. This dataset includes comments manually annotated for sentiment polarities.

The development of AI-based systems requires annotated datasets. The dataset development with subjectivity annotations are inherently subjective and influenced by annotators' standpoints, social contexts, backgrounds, etc. Additionally, political stances can affect how annotators interpret and annotate the text (Luo et al., 2020; Díaz et al., 2018). This introduces a significant gap in the emerging literature on subjectivity, particularly within the diverse Arabic linguistic context.

The implications of manual annotations in subjectivity detection are challenging, reflecting the inherently subjective nature of the task. Forms of agreement and disagreement among annotators provide insights into the subjective nature of the content and highlight the challenges in achieving consistent annotations. High agreement levels indicate clearer subjective or objective content, while disagreements reveal areas where subjectivity is more ambiguous and contested.

Such findings can show a gap in the literature on the development of subjectivity detection systems for the Arabic language. Addressing this gap re-

[3] Translated in Arabic as ذاتي ("Thati") meaning "subjective" in English.

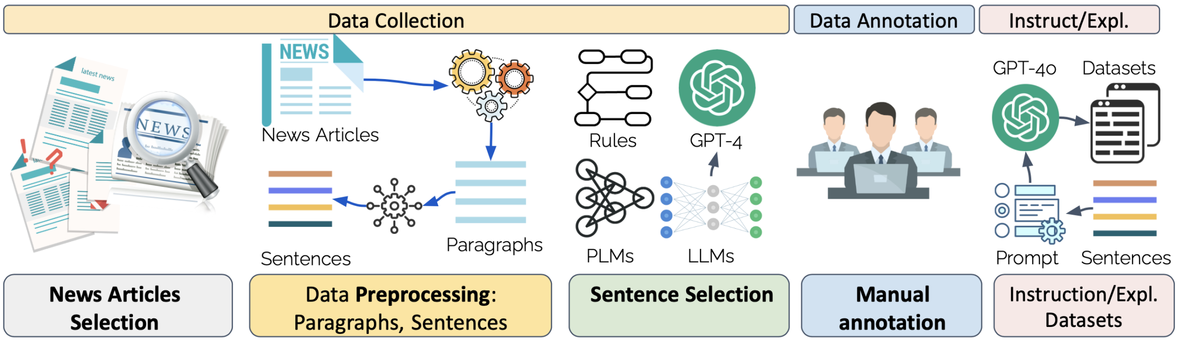

Figure 2: The pipeline of the data collection, annotation, and instruction/explanation datasets development process.

quires a deeper understanding of how these factors impact annotations and the development of more robust, context-aware AI systems.

Considering different challenges and aspects, in this study we propose *ThatiAR* dataset and provides benchmark results, which is a first of its kind for Arabic. This resource will benefit the community towards building models focusing on small and large models and conduct further resarch for news media analysis.

3 Dataset

In this section, we discuss the effort carried out to construct *ThatiAR* dataset. In Figure 2, we provide a complete workflow of the data collection (Section 3.1), manual annotation (Section 3.2) and analysis (Section 3.3).

3.1 Data Collection

To prepare a set of sentences for subjectivity annotation, we went through the following two phases.

3.1.1 News Article Selection

We selected the AraFacts dataset (Sheikh Ali et al., 2021), which contains claims verified by Arabic fact-checking websites. Each claim is associated with web pages that either propagate or refute the claim. In total, we collected 1,159 new articles from AraFacts. To address the issue of skewed distribution of sentence types in news articles, which tend to favor objective sentences, a graduate student manually searched for opinionated articles published by various Arabic news outlets (e.g., Sky News Arabia, Alarabiya). This effort resulted in selecting 221 new articles. Our pool of articles includes content from over 500 news outlets, covering a wide range of categories such as politics, social issues, arts and culture, and health, among others.

3.1.2 Preprocessing

We parsed the web pages using three different scrapers, favoring the longest output. The tools used were Goose3,[4] Newspaper3k,[5] and Trafilatura.[6] After extracting the text content, we segmented the body of text into paragraphs and sentences, resulting in a total of 15,947 sentences. The parsing and segmentation involves rule based approaches to filter and remove noisy html tags.

3.1.3 Sentence Selection

We applied two sampling strategies. The first strategy was rule-based, considering only sentences with a length between 10 and 45 words to select not overly long self-contained sentences. The second strategy utilized four transformer-based models and GPT-4 to evaluate the subjectivity of the sentences. The goal of this strategy was to select sentences with at least one "subjective" vote, thereby oversampling potentially subjective sentences for annotation. For this purpose, we trained five models: ARABERTv2 (Antoun et al., 2020), ARBERTv2 (Abdul-Mageed et al., 2021), MARBERTv2 (Abdul-Mageed et al., 2021), GIGABERTv4 (Lan et al., 2020), and GPT-4 (Achiam et al., 2023). These models were fine-tuned on the entire Arabic subjectivity datasets from the CheckThat! 2023 lab (Galassi et al., 2023). As a result, we selected 4,524 sentences, forming our annotation pool.

3.2 Data Annotation

To annotate *ThatiAR*, we opt to employ human annotators on Amazon Mechanical Turk (mTurk) crowdsourcing platform. Given that we used the mTurk platform, the demographic information of the annotators is not known to us. We adopted

[4]https://goose3.readthedocs.io
[5]https://newspaper.readthedocs.io/
[6]https://trafilatura.readthedocs.io

the annotation guidelines from a previous study (Antici et al., 2021) and tailored them for Arabic language. We discuss the annotation guidelines with additional examples in Appendix C. In a nutshell, we define *subjective* sentences as expressions of the writer's feelings, literary tastes, or personal interpretations of topics and events. Sentences containing sarcasm, support, or offensive language are also considered subjective. In contrast, *objective* sentences present facts, events, and topics based on verifiable data and include common expressions or sayings not originally written by the author.

To ensure the clarity and coherence of the guidelines, and the mTurk annotation configuration, we ran multiple pilot studies that exhausted around 850 sentences from our pool.

To ensure the quality of annotations, we sampled a set of 115 annotated sentences from CheckThat! 2023 lab (Galassi et al., 2023). We use these sentences in two ways: *(i)* 10 questions for pre-qualification test that an annotator has to pass before being eligible to start the actual HITs, and *(ii)* 105 questions for ongoing-qualification that an annotators has to maintain an acceptable accuracy throughout the annotation process. For both we requested the worker accuracy above 60%.

We finally set up the design of the annotation interface and configurations as follows. We ran 245 HITs, each containing no more than 15 sentences and 5 quiz questions. We initially collected 3 annotations per sentence and dynamically requested up to 2 more annotations when the majority agreement of 66.6% was not met, to guarantee the reliability of annotations. We compensated annotators $0.60 per HIT, costing around $550 for the entire dataset. As a result, we obtained 3,661 sentences with 66.6% agreement, of which 1,579 were subjective and 2,082 were objective sentences. The sentences that did not pass the agreement score were removed from the final dataset. In Table 2, we present a few annotated sentences from the *ThatiAR* dataset along with their English translations. Table 1 shows the statistics of *ThatiAR*.

Set	SUBJ	OBJ	All
Train	1,055 (66.8%)	1,391 (66.8%)	2,446
Dev	201 (12.7%)	266 (12.8%)	467
Test	323 (20.5%)	425 (20.4%)	748
All	**1,579**	**2,082**	**3,661**

Table 1: Statistics of *ThatiAR* dataset

3.3 Data Analysis

Annotation Agreement. To evaluate the reliability of human annotations, we computed the Inter-Annotator Agreement (IAA) using an agreement coefficient that averages the observed agreement across all annotators and sentences. We found the agreement to be approximately 0.54, indicating a reasonable level of agreement for the subjectivity annotation.

We further computed the Cohen's Kappa (C.Kappa) coefficient between each of the first three annotators and the consolidated label (determined by majority voting) (Alam et al., 2021). As shown in Table 3, the C.Kappa results indicate a moderate agreement, with an average of 0.54.[7] The annotation task for subjectivity is complex, which effects the agreement score. This complexity has also been highlighted in (Antici et al., 2021, 2024).

Deep Analysis. While manually annotating *ThatiAR*, we focused on the key aspects that impact the understanding of Arabic news reporting and the quality of annotations. We discuss our analysis by discussing the examples reported Table 7 (in Appendix).

Bias in reporting and annotating. News reports often contain phrases and terms that can be interpreted in multiple ways. Sentence #1 is example in point. The phrase "الإقليم المضطرب" ("volatile region") is a preliminary site of disagreement. The region may be described as volatile because it merits the description of Oxford dictionary definition: *"liable to change rapidly and unpredictably, especially for the worse."* However, the perception of volatility could also be influenced by partisan news reporting that portrays China as oppressive and democracy as liberating. This raises the question of whether the term "volatile" is accurate or if it carries political, historical, or cultural biases of the journalists and news agencies. On the other hand, annotators with similar potential biases are likely to consider this news sentence objective, while those with differing biases may view it as subjective.

Subjectivity in reporting and annotating. To understand the sources of disagreement between annotators, we examined several instances that ex-

[7] According to Landis and Koch's scale (Landis and Koch, 1977), Kappa values of 0.21–0.40 correspond to fair agreement, 0.41–0.60 to moderate agreement, 0.61–0.80 to substantial agreement, and 0.81–1.0 to perfect agreement.

#	Label	Sentence	Translation
1	SUBJ	وجدت بوحيرد نفسها وهي فتاة تبلغ من العمر ٢٢ عاما - بين يدي ضباط المستعمر الفرنسي ينهش لحمها بكل الطرق.	Bouhired found herself - a 22-year-old girl - in the hands of French colonial officers, a prey whose flesh was being devoured in every way.
2	SUBJ	ولكنني لم أجد الوقت الكافي للتعرف عليه عن كثب ولكن عندما مررت بأوقات عصيبة، أعطتني العقيدة الإسلامية القوة اللازمة لمواجهته.	But I did not find the time to get to know Islam closely, but when I went through difficult times, the Islamic faith gave me the strength necessary to face COVID.
3	OBJ	كما تدخل نترات الأمونيوم في صناعة المتفجرات خاصة في مجال التعدين والمناجم.	Ammonium nitrate is also used in the manufacture of explosives, especially in the field of mining.
4	OBJ	اشتية: السعوديون أعادوا القضية الفلسطينية للطاولة عند الحديث عن التطبيع مع إسرائيل	Shtayyeh: The Saudis put the Palestinian issue back on the table when talking about normalization with Israel

Table 2: Example sentences from *ThatiAR* dataset.

Setup	C.Kappa
Annotator1 vs. Majority	0.5464
Annotator2 vs. Majority	0.5512
Annotator3 vs. Majority	0.5173
Average	0.5383

Table 3: Inter-annotator agreement using Cohen Kappa (α) for *ThatiAR* dataset

hibits some aspects contributing to their subjectivity. For instance, sentence #2 references "الاحتلال" ("the occupation"), which readers commonly understand to mean "Israel." This term is politically loaded and functions as a critique of the Israeli occupation, placing blame on Israel as an occupying power and alluding to other historical occupations. Additionally, the phrase اختراق الحدود ("border breaching") followed by "الأراضي الفلسطينية" ("Palestinian lands") also carries significant political weight. If this report were from a news agency that supports Israeli claims to nationhood, it might not use "Palestinian lands" or refer to Israel as "the occupation." The term "border breaching" implies unlawful activity, indicating subjectivity in the portrayal of events. The subjectivity in this sentence may not intended as a negative or politically motivated claim but rather emerges from religious and cultural contexts that are more easily understood by regional annotators.

Composite reporting. Multiple news sentences often report different matters within the same text segment. For example, sentence #3 combines three distinct headlines into one statement, each containing both subjective and objective descriptions. This discrepancy can lead to disagreement among annotators, as each annotator may focus on different parts of the sentence or interpret the main focus differently.

Perspectives of annotators. The perspective of annotators and their standpoint is a crucial element in their judgments. For instance, a feminist annotator would account for terms in sentence #4 "قام بالتآمر" ("conspired") and " وهو عالم" ("knowing [well]"), and the usage of the term "فإن" ("if") and "فسيتخلص" ("get rid of"), therefore judge it to stand collectively as a subjective sentence. The term "conspire" implies criminal or unlawful activities; knowing, a subjective term, espouses knowledge as more prevalent in one person than another; and "get rid" implies the parents, although violent, are disposable further dehumanizing them through the lens of criminal activity. Had this statement appeared in English, the statement would appear immediately subjective. However, in the construction of the sentence in Arabic, the initial clause functions as a factual statement, further justified by the following clause, and that subjectivity may only be interpreted as appearing in the last line with the term "rid".

To this end, we answer our three questions in light of our examination and analysis of several cases and instances:

Q1. *What are the emerging characteristics of news reports with regards to subjectivity?*

Q2. *How do annotators of diverse backgrounds* *approach news reports?*

Q3. *Should current subjectivity annotation guidelines be further developed to account for morphologically-rich, socially complex, and culturally-specific content?*

To address *Q1*, we confirm that news statements either reflect an accurate description of the entities and events being reported (objective view) or convey the reporter's personal judgments and predictions about the impact of the news (subjective view). The subjective view is typically driven by political, historical, and cultural biases and subjectivities of the reporter or the news agency. We recommend hiring annotators aware of various subjectivity affecting the news being reported to ensure neutral annotations.

To address *Q2*, we affirm that annotators' political, historical, and cultural backgrounds significantly influence their understanding of the news articles and consequently their judgments. We recommend giving the annotators the option to abstain when they cannot judge sentences. This can be compiled in the annotation tool design by adding the label "Others" with the ability to provide justification, forming open-ended annotations that would be more valuable for analysis and validation.

To address *Q3*, we highlight four points:

- *Semantic Curation for Data*: We endorse the importance of carefully preparing data for annotation for ensuring accurate results. The processing pipeline, including the segmenter, must consider both syntactic and morphological aspects of the sentences. Furthermore, focusing on annotating self-contained and concise sentences will enhance the overall quality of the annotations.
- *Abstention with Open-ended Annotations*: We recommend allowing annotators to abstain when they cannot judge sentences. This is mainly because not all sentences must be subjective or objective, some are neutral or ambiguous. This can be implemented in the annotation tool by adding an "Others" label with the option to provide justification. Additionally, requesting the rationale behind annotations would enhance their value. Such open-ended annotations would be more valuable for analysis and validation.
- *Domain-specific Training for Annotation*: We emphasize the importance of specifying the data source in the annotation guidelines. For news reporting, annotators should be trained to distinguish between factual statements and text influenced by biases, as this fine distinction separates objective from subjective sentences.
- *Validation Phase for Annotation*: We highlight the importance of implementing a validation phase where annotators can meet and discuss their annotations to minimize discrepancies due to subjectivity by looking at different opinions.

4 Experimental Setup

In this section, we detail the evaluation setup used to benchmark *ThatiAR* and explore the subjectivity of Arabic news articles.

4.1 Data

We used stratified sampling to split the data into training, development, and test sets in a 70:10:20 ratio per class. Table 1 shows the statistics for each data split.

Monolingual Experiments: We used the training and development splits to fine-tune the pre-trained models. The test split was used for evaluation.

Multilingual Experiments: We used three setups for the training data: (i) AR: *ThatiAR* training set alone, (ii) ML: the entire multilingual datasets from CheckThat! 2023 (Galassi et al., 2023) and 2024 (TBA, 2024), and (iii) ALL: combining both *ThatiAR* training and the entire CheckThat! multilingual datasets. In all setups, we test on *ThatiAR* test set. We show the data statistics in Table 4.

Set	SUBJ	OBJ	All	
AR	Train	1,055 (66.8%)	1,391 (66.8%)	2,446
	Dev	201 (12.7%)	266 (12.8%)	467
	Test	323 (20.5%)	425 (20.4%)	748
ML	Train	2,580 (79.1%)	4,778 (86.0%)	7,358
	Dev	357 (11.0%)	353 (6.4%)	710
	Test	323 (9.9%)	425 (7.65%)	748
ALL	Train	3,635 (80.5%)	6,169 (85.5%)	9,804
	Dev	558 (12.4%)	619 (8.6%)	1,177
	Test	323 (7.2%)	425 (5.9%)	748

Table 4: Statistics of multilingual training data.

4.2 Models

We have used three categories of models in our experiments, dummy, pre-trained language models, and large language models.

Simple Models: To establish reasonably performing baselines, we used three simple models: RANDOM, which assigns labels randomly to sentences; MAJORITY, which assigns the most prevalent label in the dataset to all sentences; and SVC (Platt, 1998). We used standard preprocessing and TF-IDF representation to train the model using Supprt Vector Machine with its defaults parameter value to C=1.0.

Pre-trained Language Models (PLMs): We fine-tuned several PLMs to evaluate their performance on the subjectivity task using the transformer toolkit (Wolf et al., 2020).

Monolingual Experiments: We fine-tuned ARABERT version 2 (Antoun et al., 2020) and QARIB (Abdelali et al., 2021), both of which are initially trained on Arabic datasets.

Multilingual Experiments: We fine-tuned multilingual BERT (MBERT) (Devlin et al., 2019) and XLM-RoBERTa base (ROBERTA) (Conneau et al., 2020). All these models were fine-tuned using the training dataset of *ThatiAR* or the entire multilingual data from the Subjectivity Task 2 in CheckThat! Lab 2023 (Galassi et al., 2023) and 2024 (TBA, 2024).

Large Language Models (LLMs): To align with recent advancements in NLP, we experimented with Jais-13B Arabic model (Sengupta et al., 2023), GPT-4 (version 0314) (Achiam et al., 2023), Gemini-1.5 (Team et al., 2023), Mistral (Jiang et al., 2023), and Llama3-8b[8] in zero-shot setup. We also run GPT-4 in few-shot setup. For reproducibility, we set the temperature to zero for all experiments and designed the prompts using concise instructions similar to those given to human annotators when creating *ThatiAR*. We used the LLMeBench framework to run the experiments (Dalvi et al., 2024).

The use and evaluation of LLMs involve prompting and post-processing of output to extract the expected label. For each GPT-4 experimental setup we explored multiple prompts guided by the same instruction and format as recommended in in OpenAI playground. After having an expected prompt, we run complete evaluation.

Zero-Shot. For the zero-shot experiments, we designed prompts by providing natural language instructions that describe the task and specify the expected label.

[8]https://ai.meta.com/blog/meta-llama-3/

Few-Shots. For the few-shot example selection, we used the maximal marginal relevance-based (MMR) method to construct example sets that are both relevant and diverse (Carbonell and Goldstein, 1998). The MMR method calculates the similarity between a test example and the example pool (e.g., training set) and selects m examples (shots). We applied MMR on top of embeddings generated by multilingual sentence-transformers (Reimers and Gurevych, 2019). We conducted experiments with 3-shot and 5-shot examples.

4.3 Evaluation Measures

We evaluate all models' predictions using classification metrics including weighted Precision, Recall, and F1-score for the "Subjective" class.

5 Results and Discussion

5.1 Monolingual Results

Table 5 presents the benchmark results on the test set of all models trained on the *ThatiAR* training split. JAIS outperforms all other models in zero-shot setup. This highlights the importance of using models trained on Arabic data. GPT-4, in few-shot learning, markedly surpasses all other models across all measures except Recall. The optimal setup for GPT-4 is the 3-shot setup, showing a reasonable improvement compared to the 0-shot and 5-shot setups. Notably, in terms of *Recall*, JAIS, the only model trained on Arabic, outperforms GPT-4. This could indicate a weakness in GPT-4 in identifying all "Subjective" sentences, despite achieving the highest *Precision* scores by more frequently assigning the "Subjective" label to sentences.

5.2 Multilingual Results

Table 6 shows the benchmark results on the test set of all models trained on *ThatiAR* and multilingual data. The performance difference between MBERT and ROBERTA models is generally marginal across each setup. Both models achieve their best performance when fine-tuned with only Arabic data (AR setup). MBERT shows superior performance in the ALL setup, whereas ROBERTA excels in the ML setup, demonstrating its robustness in the absence of Arabic training data.

Model	Acc	P	R	F1
RANDOM	0.568	0.284	0.500	0.362
MAJORITY	0.500	0.499	0.499	0.497
SVC	0.540	0.517	0.515	0.509
QARIB	0.523	0.519	0.523	0.520
ARABERT	0.592	0.582	0.592	0.566
MBERT	0.563	0.549	0.563	0.546
ROBERTA	0.568	0.323	0.568	0.412
JAIS_{0-shot}	0.610	0.605	**0.610**	0.578
LLAMA3_{0-shot}	0.468	0.731	0.543	0.431
GEMINI_{0-shot}	0.520	0.456	0.557	0.501
MISTRAL_{0-shot}	0.539	0.167	0.238	0.415
GPT-4_{0-shot}	0.768	0.517	0.507	0.529
GPT-4_{3-shot}	**0.795**	**0.647**	0.544	**0.800**
GPT-4_{5-shot}	0.785	0.636	0.528	**0.800**

Table 5: Results of mono-lingual models on *ThatiAR*.

Setup	Model	Acc	P	R	F1
AR	MBERT	0.563	**0.549**	0.563	**0.546**
	ROBERTA	**0.568**	0.323	**0.568**	0.412
ML	MBERT	0.525	0.498	0.525	0.495
	ROBERTA	0.532	0.505	0.532	0.500
ALL	MBERT	0.554	0.535	0.554	0.528
	ROBERTA	0.532	0.502	0.532	0.494

Table 6: Results of multilingual models on *ThatiAR*. Refer to Section 4.1 for training setup, "Setup" column.

6 Annotations with Rationals

We utilized GPT-4 to validate and rationalize the human subjectivity annotations. Specifically, for each sentence in *ThatiAR*, we prompted GPT-4 with the sentence and its label, and asked it, as an expert linguist, to *"Write a simple and short explanation"* for its given annotation. We generated explanation in both Arabic and English languages, which we will release along with *ThatiAR* for the community. Table 10 (in Appendix) shows the prompt and example output in both languages.

7 Instruction Dataset

To instruct-tune LLMs, it is essential to create an instruction following dataset. For this purpose, we used GPT-4o to generate instructions for the development and test sets. To reduce the API cost of GPT-4o, the generated instructions from the development set were then used to assign instructions randomly to the samples in the training dataset. Let D_{dev} be the development set. We denote the set of instructions generated by GPT-4o for D_{dev} as I, as shown in Equation 1:

$$I = \{\text{GPT-4o}(x) \mid x \in D_{\text{dev}}\} \quad (1)$$

Let D_{train} be the training set. Instructions from I are assigned randomly to each sample in D_{train}, as represented in Equation 2:

$$\forall x \in D_{\text{train}}, \text{ assign } I_{\text{rand}}(x) \in I \quad (2)$$

where $I_{\text{rand}}(x)$ denotes an instruction randomly selected from I. This ensures that each training sample is paired with an instruction. Note that we kept the instruction from the test set independent.

To create instructions for the development and test sets, we aimed to generate diverse instructions. In Listing 1, we present the prompt used to create these instructions. For different samples, we asked GPT to create various types of instructions, such as *(i)* simple, *(ii)* straightforward, and *(iii)* detailed. We randomly selected one type from the three and used in the placeholder *random_ins_type*. The placeholder *sentence* represents the input sentence. Please see section D.2 (in Appendix) for further details.

8 Conclusion and Future Work

In this study, we propose a large subjectivity dataset for Arabic, consisting of manually annotated news sentences. We provide a detailed discussion of the data collection and annotation process. For the classification experiments, we conducted extensive experiments with PLMs and LLMs to demonstrate the utility of the dataset and system development. Additionally, we provide rationales for each sentence being classified as subjective or objective. Furthermore, we created an instruction-following dataset, which can be used in LLM-based model development.

Given the complexity of annotation, future research should include more annotators from diverse backgrounds to further enhance the subjectivity annotation process. Our study is preliminary in nature and serves as an initial step towards understanding news media in terms of subjectivity. However, this study presents important considerations for scholars specifically interested in subjectivity and for the field of NLP in general.

9 Limitations

Subjectivity annotation is a complex task, which has also been noticed in other languages. Even though we provided clear guideline in Arabic to make sure that native speakers fully understand the task, however, it still become a challenge for that. Many mturk annotators did not pass our qualified test. It might be because they are from diverse background, and culture, which might have effected the annotation process.

Ethics and Broader Impact

We collected news articles from a range of Arabic media outlets and selected sentences for annotation. While we aimed to include diverse topics and perspectives, we acknowledge the potential for bias in our data sampling. Annotations are inherently subjective and may reflect the sociocultural biases of the annotators. To mitigate this, we recruited annotators from different Arabic-speaking countries, with diverse educational and professional backgrounds. We also developed detailed annotation guidelines and conducted multiple rounds of training to promote consistency. However, biases and disagreements remain, which we analyze in the discussion section. In any of the data collection and annotation process we do not collect any personally identifiable information.

The models developed using ThatiAR have significant potential for positive impact by helping to detect subjective and potentially biased or misleading content in Arabic news. This can assist fact-checkers, journalists, and policymakers in combating misinformation and promoting media literacy. However, we also recognize the potential for misuse, such as in censorship or political manipulation. We encourage users to consider the ethical implications of their applications. Furthermore, while ThatiAR is a step towards greater representation of Arabic in NLP research, much work remains to fully capture the linguistic diversity of Arabic and §its dialects. Our annotators and data sources skew towards Modern Standard Arabic, which may not reflect everyday language use. Future work should prioritize inclusivity and linguistic diversity. We are releasing ThatiAR dataset and resources publicly to encourage research on Arabic subjectivity analysis. However, we urge researchers to be transparent about the limitations and potential biases of the dataset and any resulting models. Appropriate documentation should be provided to help end users make informed decisions about model deployment.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Abdul-Mageed. 2015. *Subjectivity and sentiment analysis of Arabic as a morophologically-rich language*. Ph.D. thesis, Indiana University.

Muhammad Abdul-Mageed and Mona Diab. 2011. Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th linguistic annotation workshop*, pages 110–118.

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. SAMAR: subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Muhammad Abdul-Mageed and Mona T Diab. 2012. AWATIF: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, volume 515, pages 3907–3914.

Muhammad Abdul-Mageed and Mona T Diab. 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Mohamed Elhag M Abo, Nordiana Ahmad Kharman Shah, Vimala Balakrishnan, Mohamed Kamal, Ahmed Abdelaziz, and Khalid Haruna. 2019. Ssasda: subjectivity and sentiment analysis of sudanese dialect arabic. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5. IEEE.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Francesco Antici, Luca Bolognini, Matteo Antonio Inajetovic, Bogdan Ivasiuk, Andrea Galassi, and Federico Ruggeri. 2021. Subjectivita: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 40–52, Cham. Springer International Publishing.

Francesco Antici, Federico Ruggeri, Andrea Galassi, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2024. A corpus for sentence-level subjectivity detection on English news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 273–285, Torino, Italia. ELRA and ICCL.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Aqil M Azmi and Samah M Alzanin. 2014. Aara'–a system for mining the polarity of saudi public opinion through e-newspaper comments. *Journal of Information Science*, 40(3):398–410.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675—-718, Indonesia. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, and Ahmed Ali. 2024. LLMeBench: A flexible framework for accelerating LLMs benchmarking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 214–222, St. Julians, Malta. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Andrea Galassi, Federico Ruggeri, Alberto Barrón-Cede no, Firoj Alam, Tommaso Caselli, Mucahid Kutlu, Julia Maria Struss, Francesco Antici, Maram Hasanain, Juliane Köhler, Katerina Korre, Folkert Leistra, Arianna Muti, Melanie Siegel, Mehmet Deniz Turkmen, Michael Wiegand, and Wajdi Zaghouani. 2023. Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles. In *Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum*, CLEF '2023, Thessaloniki, Greece.

Nizar Habash, Behrang Mohit, Ossama Obeid, Kemal Oflazer, Nadi Tomeh, and Wajdi Zaghouani. 2013. Qalb: Qatar arabic language bank. In *Proceedings of Qatar Annual Research Conference (ARC-2013), pages ICTP–032, Doha, Qatar*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Hairong Huo and Mizuho Iwaihara. 2020. Utilizing bert pretrained models with various fine-tune methods for subjectivity detection. In *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*, pages 270–284. Springer.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Mohammed Korayem, David Crandall, and Muhammad Abdul-Mageed. 2012. Subjectivity and sentiment analysis of arabic: A survey. In *Advanced Machine Learning Technologies and Applications: First International Conference, AMLTA 2012, Cairo, Egypt, December 8-10, 2012. Proceedings 1*, pages 128–139. Springer.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. Gigabert: Zero-shot transfer learning from english to arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64.

John Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.

Eshrag Refaee and Verena Rieser. 2014a. An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.

Eshrag Refaee and Verena Rieser. 2014b. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, volume 16.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 305–314, Toronto, Canada. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

TBA. 2024. Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles. In *Working Notes of CLEF 2024–Conference and Labs of the Evaluation Forum*, CLEF '2024, Grenoble - France.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Data Release

The *ThatiAR* dataset[9] is released under the Creative Commons Attribution 4.0 International License: https://creativecommons.org/licenses/by/4.0/legalcode. The dataset includes the following files:

- Subjectivity manual annotations divided into training, development, and test sets, in CSV format. Each news sentence is represented by an id, text, and label.
- Annotation guidelines provided to the crowd annotators in Arabic.
- Explanation and instruction annotations generated automatically by the GPT-4o model, in JSONL format, with the same splits as the manually annotated data. Each news sentence is represented by an id, text, label, explanation, and instruction.
- Example scripts for running experiments, including PLMs (AraBERT model) and LLMs (GLUE model).

B Details of the experiments

For the experiments, we used SVM, PLMs, and LLMs (GPT-4). All these scientific artifacts are used according to their terms and conditions for research purposes. Below, we discuss the parameters we used. Furthermore, we have made all our scripts available to ensure reproducibility.

Models and Parameters:

- **AraBERT**: L=12, H=768, A=12; the total number of parameters is 371M, where *L* is the number of layers (i.e., Transformer blocks), *H* is the hidden size, and *A* is the number of self-attention heads;
- **BERT Multilingual** (bert-base-multilingual-uncased) (mBERT): L=12, H=768, A=12, number of parameters (172M);
- **XLM-RoBERTa** (xlm-roberta-base): L=24, H=1027, A=16; the total number of parameters is 355M.
- N-gram with SVM: TF-IDF transformation and used C=1.0 in SVM.

To fine-tune PLMs, we used the following hyper-parameters.

- Batch size: 8;
- Learning rate (Adam): 2e-5;
- Number of epochs: 10;
- Max seq length: 256.

We ran the PLM-based fine-tuning experiments with different seed values and report the results of the best runs on the development set of *ThatiAR*. We run our experiments on a cluster consisting of GPUs such as P100, V100, V100-NVLINK, and T4.

C Annotation Guidelines

For the annotation we adopted and refined the annotation guidelines discussed in (Antici et al., 2021). To begin the annotations, annotators of diverse backgrounds were provided with a specific use-cases for subjective and objective sentences that we present in Tables 7, 8 and 9, respectively. We release the annotation guidelines with the dataset.[10]

C.1 Subjective Use Cases

We define *subjective* sentences as expressions of feelings, literary tastes, or personal interpretations of topics and events. Below are a few use cases of subjective sentence with examples in Table 8:

- Sentences expressing personal opinions about events and topics, or containing rhetorical questions, or containing probabilities and expectations and building conclusions on them, e.g., Sentence #1.
- Sentences containing sarcasm or humor, according to the writer's expression, e.g., Sentence #2.
- Sentences encouraging, supporting, or approving an action , e.g., Sentence #3.
- Sentences containing offensive expressions such as racism, tactlessness, etc., e.g., Sentence #4.
- Sentences containing a rhetorical expression and depiction of people and situations, such as "exaggeration", that a writer uses to express his or her personal opinion, e.g., Sentence #5.

C.2 Objective Use Cases

We define *objective* sentences as a presentation of facts, events, and topics based on factual data. Below are a few use cases of objective sentence with examples in Table 9:

- Sentences containing news (Sentence #1), facts (Sentence #2) and laws (Sentence #3) conveyed by the writer of the sentence.

[9] anonymous.com

[10] anonymous.com

#	Sentence	Translation
1	دمرت السلطات الصينية آلاف المساجد في شينجيانغ، حسبما ذكر مركز أبحاث أسترالي الجمعة، في أحدث تقرير عن انتهاكات واسعة لحقوق الإنسان في الإقليم المضطرب.	Chinese authorities have destroyed thousands of mosques in Xinjiang, an Australian think tank said Friday, in the latest report on widespread human rights violations in the volatile region.
2	وأعلن الاحتلال في أوقات المسيرات خلال الأعوام السابقة عن رفع حالة الاستنفار العسكري على الحدود لأكثر من أسبوعين، خشية اختراق الحدود من دول الطوق باتجاه الأراضي الفلسطينية، والذين من ضمنهم مئات من المتضامنين الدوليين.	The occupation announced during marches throughout recent years in reducing military presence in borders for more than two weeks for fear of breaching the borders from the countries of the enclave towards the Palestinian territories, including hundreds of international solidarity activists.
3	هالة صدقي تلجأ إلى الديانة الإسلامية لتحقيق رغبتها وهذه التفاصيل ـ مشاهير عالمية كم تقاضى جونقكوك من فرقة BTS للغناء في إفتتاح مونديال قطر؟ ـ المشاهير العرب مفاجأة من العيار الثقيل!!	Hala Sudqi resorts to the Islamic religion to achieve her desires and these are the details – celebrities globally how much does Jongkok from BTS make for singing in the opening ceremony of FIFA in Qatar? – Arabic celebrities, a big surprise!
4	وتبين أيضا أن زوجة الجاني سبق ان قامت بقطع المساعدة المالية عن ابنهما، وأن هذا الأخير قام بالتآمر على والديه عن طريق حشو المسدس بالرصاص، وهو عالم بما دأب عليه أبوه من عادة تهديد أمه بالقتل عن طريق ذلك المسدس الفارغ، فإن نفذ تهديده مرة واحدة فسيتخلص من أمه وأبيه بضربة، أو رصاصة واحدة.	It was also found that the perpetrator's wife had previously cut off financial assistance from their son, and the son had conspired against his parents by loading the gun knowing that his father usually threatened his mother with death using an empty gun. If the father executes his threat one more time, then the son will be rid of both his parents at once or should we say with one bullet.

Table 7: Examples of news sentences.

- Sentences describing the writer's feelings or emotions without expressing any opinions on any topic, e.g., Sentence #4.
- Sentences containing opinions, claims, feelings, or viewpoints attributed to a third party other than the writer, e.g., Sentence #5.
- Sentences conveying the writer's comments without explicitly stating any personal conclusion, interpretation, or expression of a personal opinion, so that the discussion is left open, e.g., Sentence #6.
- Sentences stating conclusions reached by the writer of the sentence, without expressing his personal position or opinion, or they are justified by hypotheses that are not related to personal opinions, e.g., Sentence #7.
- Sentences referring to an individual by a well-known nickname that was not given by the writer, e.g., Sentence #8.
- Common expressions and examples or sayings, e.g., Sentence #8.

D Challenges

D.1 Annotation Challenges

Annotating for subjectivity presents significant challenges, especially when conducted via crowdsourcing platforms. One major obstacle is the lack of shared cultural, linguistic, and experiential backgrounds among annotators. As mTurk does

#	Sentence	Translation
1	والدليل اغتيال البروفيسور التونسي محمود عبد القادر البزرتي وكان قد فك الشيفرة الوراثية للفيروس وعزلها مخبريا واعادة تشكيلها ليتحول الفيروس الى لقاح مضاد وذلك أغضب أمريكا فقامت باغتياله إنها باختصار جريمة بحق البشرية.	The evidence is the assassination of Tunisian professor Mahmoud Abdel Qader Al-Bazrti, who had decoded the genetic code of the virus, isolated it in a laboratory, and reconstructed it so that the virus could be transformed into an anti-vaccine. This angered America, so it assassinated him. It is, in short, a crime against humanity.
2	لنعود كل سنة مع أول قطرة غيث إلى نقطة الصفر نشكو لهم من انسداد البالوعات فيقولوا لنا عليكم بالسباحة!!!!	Let us return back every year with the first drop of rain to ground zero and complain to them about the clogged drains, and they tell us that you should go swimming!!!!
3	هذا يعني حماقة أو بيستحمق الآخرين، فنحن نعتقد من واجبنا شرعاً أن نقاوم هذا الاحتلال بكل ما أوتينا من قوة ونعاقبه بنفس الطرق التي هو يستخدمها ضدنا.	This means foolishness or fooling others. We believe it is our legal duty to resist this occupation with all our might and punish it with the same methods it uses against us.
4	كما سقط مبارك وسيسقط السفيه السيسي، وخطيئة أى ديكتاتور أن ينظر إلى الشعب من خلال عصابته المنتفعة، وأن يصدق نفاقهم ويكذب علامات الغضب الثوري.	Mubarak also fell and the foolish Sisi will fall, and it is the sin of any dictator to look at the people through his beneficial gang, believe their hypocrisy and deny the signs of revolutionary anger.
5	لا تظلم أحداً، فالظلم نار لا تنطفئ في قلب صاحبها، ولو مرّت عليه الأعوام.	Do not oppress anyone, for injustice is a fire that will never be extinguished in the heart of its perpetrator, even if years have passed.

Table 8: Use cases of "subjective" sentences.

not disclose demographic information about annotators, we did not analyze the effect of educational, cultural, and regional backgrounds on annotation. Such disparities inevitably influence how annotators interpreted sentences and judged subjectivity, leading to disagreements.

Furthermore, subjective annotations are inherently influenced by individual biases, standpoints, and opinions, which are difficult to control in a crowdsourced setting. Achieving reliable annotations required iterative refinement of guidelines, pilot studies, qualifications tests, and ongoing quality checks – underscoring the complexities of crowdsourcing subjective annotations across diverse annotator pools.

D.2 Prompting Challenges

The performance of the model is highly dependent on the prompting strategy. Designing optimal prompts for each task is challenging and requires multiple iterations. Depending on the prompt, the output varies across all instances of the dataset. For the subjectivity task in this study, we experimented with *(i)* zero-shot and few-shot methods for label generation, *(ii)* generating explanations, and *(iii)* generating the instruction dataset. In Table 10, we provide examples of prompts in Arabic and English for generating explanations, along with the provided sentences and their labels.

```
prompt = f"You are an expert in creating
    instruction datasets to train AI
    models. \
Here, our idea is to create an
    instruction dataset for a
    subjectivity detection task. \
The task is to determine whether a
    sentence is subjective or objective.
    \
Write a '{random_ins_type}' instruction
    for this '{sentence}'. Do not
    include the sentence in the
    instruction."
```

Listing 1: Prompt to create instructions.

#	Sentence	Translation
1	و أضافت نفس المصادر أن هناك أحزاب رفضت هذا الاقتراح نظرا لأن بن عيسى غير ملم بالوضع الاقتصادي الحرج التي تمر به تونس.	The same sources added that there are parties that rejected this proposal because Ben Aissa is not familiar with the critical economic situation that Tunisia is going through.
2	ويمثل الفيروس خطرا بشكل خاص على كبار السن وعلى مَن يعانون من مشاكل صحية، ولدى إيطاليا واحدة من أكبر نسب كبار السن في العالم.	The virus represents a particular danger to the elderly and those suffering from health problems, and Italy has one of the largest proportions of elderly people in the world.
3	في حالة وجود منظمة غير حكومية مرخص لها بتصدير أو إعادة تصدير الخدمات لسوريا بموجب هذا القسم في التاريخ السابق لتاريخ سن هذا القانون، فإن هذا القسم ينطبق على هذه المنظمة في تاريخ سن القانون وبعده إلى الحد نفسه وبنفس الطريقة التي كان ينطبق بها هذا القسم على هذه المنظمة في التاريخ السابق لتاريخ سن هذا القانون.	If a non-governmental organization licensed to export or re-export services to Syria under this section existed on the date prior to the date of enactment of this Act, this section shall apply to such organization on and after the date of enactment of this Act to the same extent and in the same manner as this section applied to such The organization on the date prior to the date of enactment of this law.
4	وكلما عدت من التحقيق والمحاكم منهكة إلى زنزانتي أجدد عهدي لله بأنني رغم كل التضييقات سأسير حتى مماتي على نفس الدرب لأنني نذرت نفسي للمسجد الأقصى.	Whenever I return from the investigation and the courts exhausted to my cell, I renew my pledge to God that despite all the restrictions, I will walk the same path until my death because I have vowed myself to Al-Aqsa Mosque.
5	وتقول إنه في المجتمع العلمي، هناك تعاريف متضاربة حول موعد تحقيق مناعة القطيع.	She says that in the scientific community, there are conflicting definitions about when herd immunity will be achieved.
6	لقد كانت جميلة اسما على مسمى، انطلقت بأحلام الزهور فتعلمت الخياطة والرقص وكانت تحلم بأن تكون مصممة أزياء، ولكن القدر قادها لتكون مصممة أجمل نضال في تاريخ البشرية.	She was beautiful by her name. She started out with dreams of flowers, learned sewing and dancing, and dreamed of being a fashion designer, but fate led her to become the designer of the most beautiful struggle in human history.
7	في حال تعرض الفيروس لدرجة حرارة ٦٢ أو ٧٢ سوف يُقتل لذا لا يعيش في المناطق الحارة.	If the virus is exposed to a temperature of 26 or 27, it will be killed, so it does not live in hot areas.
8	وجاء ليصب في مصلحة القوى اليمينية الرافضة لمفهوم الوحدة، مثل تيار لوبان في فرنسا، أو حزب الاستقلال البريطاني، غير أن رافضي الدستور الفرنسيين لم يكونوا جميعاً من التيار اليميني بل كانوا أيضاً من التيار اليساري، وسبب رفضهم يرجع إلى انتقادهم للسياسة الاقتصادية المقترحة التي لا تهتم بتأمين ضمانات اجتماعية كافية.	It came to serve the interest of the right-wing forces that reject the concept of unity, such as the Le Pen movement in France, or the British Independence Party. However, the French opponents of the constitution were not all from the right-wing movement, but rather they were also from the left-wing movement, and the reason for their rejection is due to their criticism of the proposed economic policy that does not care about Providing adequate social guarantees.
9	الأدعية المأثورة: ""اللهم إني أعوذ بك من البرص والجنون والجذام وسيئ الأسقام وأعوذ بك من همزات الشياطين وأعوذ بك رب أن يحضرون وصلِ اللهم على سيدنا محمد وآل محمد""	Supplications: "O God, I seek refuge in You from leprosy, madness, leprosy, and bad diseases, and I seek refuge in You from the instigations of the devils, and I seek refuge in You, my Lord, from their presence, and may God bless our master Muhammad and his family."

Table 9: Use cases of "objective" sentences.

Input Prompt	أنت خبير لغوي. تستطيع فهم التعبير الذاتي في جمل المقالات الإخبارية. اشرح لماذا الجملة التالية: «ومع أن أداء القوات المصرية كان نصرًا لا يمكن لمنصف أن يماري فيه، فإن نظام أنور السادات كان امتدادًا للنظام العربي الرسمي، الذي استولدَه الاستعمار الغربي، وجعله بمثابة غشاوة استقلال خادعة تعمي عيون وبصائر الشعوب العربية، وتطمس على قلوبها فتغفل عن واقعها المرير». تعتبر «ذاتية». اكتب شرح موجز وبسيط.
GPT Response	الجملة المذكورة تعتبر ذاتية لأنها تحتوي على تقييم شخصي لأداء القوات المصرية ونظام أنور السادات. الكاتب يعبر عن رأيه الشخصي بأن أداء القوات المصرية كان نصرًا لا يمكن الجدال فيه، ولكنه ينتقد نظام السادات بوصفه امتدادًا للنظام العربي الذي أنشأه الاستعمار الغربي. الكاتب يستخدم عبارات مثل «غشاوة استقلال خادع» و«تعمي عيون وبصائر الشعوب العربية» للتعبير عن وجهة نظره الشخصية حول النظام السياسي في مصر والعالم العربي. هذه العبارات تظهر تحيز الكاتب وتقييمه الشخصي للأحداث، مما يجعل الجملة ذاتية.
Input Prompt	You are an expert linguist. You can understand the subjectivity of sentences in news articles. Explain why the sentence is "Subjective": «ومع أن أداء القوات المصرية كان نصرًا لا يمكن لمنصف أن يماري فيه، فإن نظام أنور السادات كان امتدادًا للنظام العربي الرسمي، الذي استولدَه الاستعمار الغربي، وجعله بمثابة غشاوة استقلال خادعة تعمي عيون وبصائر الشعوب العربية، وتطمس على قلوبها فتغفل عن واقعها المرير». Write a simple and short explanation.
GPT Response	The sentence in question contains subjective language because it includes value judgments and opinions rather than objective facts. Phrases like "نصرًا لا يمكن لمنصف أن يماري فيه" (a victory that no fair person can dispute) and "غشاوة استقلال خادعة" (a deceptive veil of independence) reflect the author's personal perspective and interpretation of events. Additionally, the use of emotionally charged words such as "تعمي" (blind) and "تطمس" (obscure) further indicates a subjective viewpoint. The sentence presents the author's opinion on the performance of the Egyptian forces and the nature of Anwar Sadat's regime in relation to Western colonialism and its impact on Arab societies.

Table 10: Prompt for generating explanations for human-annotated sentences